Il genoma umano

Le meraviglie della complessità
Presupposti, implicazioni, effetti

Clelia Peano, 11 Novembre, 2026

Slide 01 Il genoma umano

Oggi vorrei portarvi dentro qualcosa che ciascuno di noi possiede, ma che nessuno può vedere a occhio nudo: il genoma umano. È l'insieme del nostro DNA e contiene informazioni che contribuiscono allo sviluppo e al funzionamento dell'organismo. Lo esploreremo attraverso il tema di questa conferenza, la complessità. Vedremo quali sono i suoi presupposti, che cosa implica studiarli e quali effetti hanno le nostre scoperte. Il punto da tenere presente fin dall'inizio è che il DNA agisce dentro cellule, organismi e ambienti. Per capire una persona dobbiamo comprendere queste relazioni.

Slide 02 Da tre miliardi di lettere a una persona

Partiamo da questa domanda: come fanno circa tre miliardi di lettere a contribuire alla formazione di una persona? Le lettere del DNA sono soltanto quattro, A, C, G e T. A cambiare sono il loro ordine e il modo in cui la cellula usa le diverse sequenze. Pensate a quante parole possiamo costruire con un alfabeto limitato. Nel genoma, però, oltre al testo contano i processi che lo leggono e lo regolano. La domanda quindi si allarga: come viene utilizzata questa informazione, durante lo sviluppo e nel corso della vita?

Slide 03 La complessità in tre passi

Queste tre parole saranno la nostra bussola. Con presupposti intendo le caratteristiche del genoma da cui partiamo: la sequenza, i geni, le ripetizioni e la sua organizzazione. Le implicazioni riguardano ciò che serve per comprenderlo: tecnologie capaci di leggerlo e conoscenze su come viene regolato. Gli effetti sono ciò che possiamo ottenere, dalle diagnosi alle terapie, fino alla ricostruzione della nostra storia evolutiva. I tre aspetti sono collegati: una nuova tecnologia permette una scoperta, e quella scoperta può suggerire una nuova domanda.

Slide 04 Che cos’è la complessità

Guardate lo stormo. Ogni uccello interagisce con quelli vicini e da queste interazioni emerge una forma collettiva. Questa è una buona immagine per introdurre la complessità: molte componenti possono produrre comportamenti che non riconosciamo osservandone una sola. Nel genoma incontreremo qualcosa di simile, pur con meccanismi completamente diversi. Una sequenza può avere un effetto diverso a seconda della cellula, delle altre sequenze e dell’ambiente. Per questo contare le parti è utile, ma dobbiamo anche studiare come interagiscono.

Slide 05 Come si impacchettano due metri di DNA

Seguiamo ora il percorso dal piccolo al grande. I nucleotidi sono le unità del DNA e la doppia elica è la struttura che li organizza. Il DNA si avvolge intorno a proteine chiamate istoni, formando i nucleosomi. DNA e proteine costituiscono la cromatina, che si ripiega e si organizza nello spazio. In una tipica cellula umana diploide ci sono circa due metri di DNA, contenuti in un nucleo grande pochi micrometri. Questo impacchettamento è dinamico: aiuta anche a rendere alcune regioni più o meno accessibili. Il cromosoma molto condensato è particolarmente evidente durante la divisione cellulare.

Slide 06 I numeri del genoma umano

Ecco tre numeri che ci aiutano a orientarci. Un corredo di cromosomi contiene circa 3,1 miliardi di paia di basi. Le tipiche cellule del nostro corpo possiedono due corredi, uno ricevuto da ciascun genitore: in totale 23 coppie di cromosomi. Nel genoma si trovano circa ventimila geni che codificano proteine. Questi numeri descrivono cose diverse: la lunghezza del testo, il numero dei suoi grandi volumi e alcune delle informazioni che contiene. Esistono eccezioni cellulari, ma questa è la struttura di riferimento da cui partire.

Slide 07 La maggior parte del DNA non codifica proteine

La parte che codifica proteine occupa soltanto circa l'uno o due per cento del genoma. Nel resto troviamo introni, sequenze regolatrici, geni per RNA non codificanti e moltissime ripetizioni. Alcune sequenze ripetute derivano da elementi capaci, nella storia evolutiva, di spostarsi o moltiplicarsi nel genoma. Possiamo quindi immaginare un testo che contiene istruzioni, segnali di controllo, frammenti ripetuti e tracce della propria storia. Per una parte di queste sequenze conosciamo funzioni precise; per altre stiamo ancora cercando di capire se e come contribuiscano al funzionamento delle cellule.

Slide 08 Più DNA non significa più complessità

Potremmo pensare che un organismo più complesso abbia necessariamente un genoma più grande. Questo confronto mostra che non funziona così. La cipolla ha molto più DNA di noi e l'axolotl, un anfibio, ne possiede ancora di più. Il cosiddetto paradosso del valore C riguarda proprio questa relazione poco intuitiva tra quantità di DNA e complessità. Le dimensioni possono aumentare per duplicazioni e accumulo di ripetizioni. Anche il numero dei geni, da solo, non permette di classificare la complessità degli organismi. Per capire le differenze dobbiamo studiare organizzazione, regolazione e storia evolutiva.

Durante l'esposizione *Chiedi quale specie gli studenti avrebbero previsto in cima al grafico.*

Slide 09 Un gene più messaggi possibili -SPLICING ALTERNATIVO

Anche dentro un singolo gene possiamo trovare più possibilità. Quando viene prodotto un RNA iniziale, alcune sue parti, gli introni, vengono rimosse; altre, gli esoni, vengono unite. Lo splicing alternativo permette di combinare gli esoni in modi diversi. Nello schema, un messaggio conserva tutti gli esoni rappresentati e un altro ne salta uno. Da uno stesso gene possono così derivare RNA differenti e, in molti casi, forme diverse di una proteina. Il risultato dipende anche dal tipo cellulare e dalla situazione. Non ogni RNA alternativo produce però una proteina stabile.

Slide 10 Un testo enorme ripetitivo e organizzato a più livelli

Fermiamoci sui presupposti della complessità. Abbiamo un testo molto lungo, pieno di ripetizioni, organizzato nello spazio e capace di produrre messaggi diversi. Queste caratteristiche spiegano anche perché leggerlo integralmente sia stato difficile. Immaginate di dover ricostruire un libro da tanti frammenti: nelle pagine con frasi ripetute, dove mettereste ciascun pezzo? Questa difficoltà ci porta alla storia del sequenziamento del genoma umano.

Slide 11 Il primo grande atlante Human Genome Project

Il Progetto Genoma Umano iniziò nel 1990 e si concluse nel 2003, dopo circa tredici anni. Coinvolse una collaborazione internazionale e produsse una risorsa di riferimento per tutta la ricerca. La cifra di circa 2,7 miliardi di dollari riguarda il programma statunitense nel suo complesso, non il costo di un singolo test. Fu un lavoro enorme anche sul piano della condivisione dei dati e dello sviluppo dei metodi. Quel riferimento rese possibili molti confronti e nuove scoperte. La conclusione del progetto, però, lasciava ancora alcune regioni del genoma difficili da ricostruire.

Slide 12 Perché restavano dei buchi

I buchi rimasti erano concentrati soprattutto nelle regioni più ripetitive. Fra queste troviamo i centromeri, importanti per la distribuzione dei cromosomi durante la divisione, e tratti di DNA satellite o duplicazioni molto simili tra loro. Se una breve sequenza compare in molti punti, un frammento può essere letto correttamente ma risultare difficile da collocare. È come trovare un pezzetto di pagina con una frase presente in tanti capitoli. Restava da ricostruire circa l’otto per cento della sequenza, comprendente regioni che potevano contenere informazioni biologiche importanti.

Slide 13 T2T un genoma senza lacune nel 2022

Nel 2022 il consorzio Telomere-to-Telomere pubblicò T2T-CHM13. Il nome indica l'obiettivo di ricostruire i cromosomi da un'estremità all'altra, senza le precedenti lacune. Furono aggiunte circa duecento milioni di basi. Il risultato combinava letture lunghe PacBio HiFi e letture Nanopore molto lunghe, usando un campione in cui le due copie dei cromosomi erano quasi identiche e quindi più facili da ricostruire. Quella sequenza comprendeva 22 autosomi e il cromosoma X. Il cromosoma Y fu completato separatamente nel 2023. Era un riferimento completo per quel campione, mentre la diversità umana richiede molti altri genomi.

Slide 14 Il vantaggio vedere dove prima non si riusciva

Perché colmare quei buchi è importante? Perché il riferimento è la mappa sulla quale collochiamo le sequenze di un nuovo campione. Se una regione manca o è rappresentata male, alcune letture possono essere scartate o attribuite al punto sbagliato. Un riferimento più completo permette di studiare meglio ripetizioni, centromeri e copie geniche molto simili. Alcune differenze diventano quindi più facili da localizzare e confrontare. Lo schema mostra questo principio: la variante non compare dal nulla, ma diventa osservabile con una mappa migliore. Capire quale effetto abbia richiede poi ulteriori studi.

Slide 15 Stesso DNA cellule molto diverse

Guardate queste cellule: un neurone, una cellula muscolare e una cellula immunitaria svolgono lavori molto diversi. Eppure condividono in gran parte lo stesso patrimonio genetico. La differenza nasce soprattutto da quali geni vengono utilizzati e con quale intensità. Un neurone deve produrre componenti per trasmettere segnali; una cellula muscolare deve sostenere la contrazione. Il genoma contiene molte possibilità, ma ogni tipo cellulare ne utilizza una combinazione particolare. Entriamo così nella regolazione dell'espressione genica: il livello che aiuta a spiegare come uno stesso testo contribuisca a funzioni differenti.

Slide 16 Un interruttore può agire da molto lontano

Un gene viene trascritto a partire da una regione chiamata promotore. Altre sequenze, gli enhancer, possono contribuire a regolarne l'attività anche da molto lontano lungo il DNA. Questo è possibile perché il DNA si ripiega: regioni separate da migliaia o centinaia di migliaia di basi possono avvicinarsi nello spazio. Le proteine regolatrici aiutano queste interazioni. Guardando soltanto una sequenza lineare, rischieremmo quindi di non riconoscere una relazione importante. Una variante in un enhancer può cambiare quanto viene espresso un gene anche lasciando identica la sequenza della proteina.

Slide 17 Epigenetica annotazioni sul testo

Alla sequenza si aggiungono modifiche chimiche che influenzano il contesto nel quale il DNA viene utilizzato. La metilazione riguarda il DNA; altre modifiche interessano gli istoni, le proteine intorno alle quali il DNA si avvolge. La metafora dei post-it è utile: le lettere restano le stesse, ma troviamo annotazioni che aiutano a organizzarne l'uso. L'effetto, però, dipende dalla posizione della modifica e dal

tipo di cellula. Non esiste una regola universale per cui una modifica accende sempre un gene e un'altra lo spegne. Anche qui servono contesto ed esperimenti.

Slide 18 Il genoma ha anche un'architettura 3D

Quando pensiamo al DNA, spesso immaginiamo una lunga doppia elica, come una riga di testo. Ma dentro il nucleo il DNA è ripiegato nello spazio, e il modo in cui si organizza contribuisce a regolare l'attività dei geni. Questa immagine ci accompagna attraverso tre livelli di ingrandimento. Nel primo vediamo i territori cromosomici. Ogni cromosoma occupa prevalentemente una regione del nucleo, chiamata "territorio". Possiamo immaginare il nucleo come una città e i territori come i suoi quartieri. I confini, però, non sono pareti: le regioni possono entrare in contatto e la loro organizzazione è dinamica. Nel secondo ingrandimento osserviamo la cromatina, cioè il DNA associato a proteine. Il DNA si avvolge intorno agli istoni e si ripiega formando anse e domini: regioni nelle quali alcuni tratti tendono a incontrarsi più frequentemente. Questo ripiegamento non serve soltanto a far entrare il DNA nel nucleo. Contribuisce anche a rendere alcune regioni più accessibili e a favorire determinati incontri tra sequenze. Ed eccoci al terzo ingrandimento: l'interazione tra enhancer e promotore. Il promotore è una regione di DNA vicina al punto in cui inizia la trascrizione di un gene. Qui si riuniscono le proteine necessarie per avviare la produzione di RNA. L'enhancer è invece una regione regolatrice che può trovarsi molto lontano lungo la sequenza del DNA. Quando viene riconosciuto dalle proteine appropriate, può contribuire ad aumentare la trascrizione di un gene. Come fanno due regioni lontane a comunicare? Il DNA si ripiega: formando un'ansa, enhancer e promotore possono avvicinarsi nello spazio. Le proteine legate a queste regioni possono così interagire e favorire l'attività della macchina che produce l'RNA. Pensiamo a un lungo nastro con due segni distanti. Se lo distendiamo sul tavolo, i segni sono lontani; se lo pieghiamo, possono trovarsi uno accanto all'altro. La distanza lungo il nastro non coincide con la distanza nello spazio. L'incontro, però, non garantisce da solo che un gene venga attivato: contano anche le proteine presenti, lo stato della cromatina e i segnali ricevuti dalla cellula. Questo aiuta a capire perché un neurone e una cellula del fegato, pur possedendo sostanzialmente lo stesso DNA, utilizzino programmi diversi di attività genica. L'organizzazione tridimensionale è uno dei livelli che contribuiscono a queste differenze. La frase da ricordare è: "Lontani lungo il DNA, vicini nello spazio". La complessità del genoma nasce quindi anche dalle relazioni tra le sue parti: non conta soltanto quali sequenze possediamo, ma anche come possono incontrarsi e lavorare insieme

Slide 19 e 20 Due esempi la regolazione cambia la vita

Per capire il genoma dobbiamo conoscere le istruzioni contenute nei geni, ma anche capire come le cellule le utilizzano. Quanto deve lavorare un gene? In quali cellule? In quale momento della vita?

Gli **enhancer** contribuiscono a questo controllo: sono tratti di DNA che possono aumentare l'attività di un gene. Possiamo immaginarli come regolatori del volume. Le proteine che si legano a queste regioni aiutano ad aumentare la produzione di un determinato messaggio.

Nella parte sinistra vediamo un esempio che riguarda l'alimentazione.

Il gene *LCT* contiene le istruzioni per produrre la **lattasi**, l'enzima che permette di digerire il lattosio, lo zucchero del latte. In molte persone la produzione di lattasi diminuisce dopo l'infanzia. In altre rimane elevata anche in età adulta.

Una delle spiegazioni si trova in un enhancer situato dentro un introne del gene vicino, *MCM6*. Questo è già un aspetto interessante della complessità: una regione dentro un gene può contribuire a regolarne un altro.

Alcune varianti di questo enhancer, come quella indicata nella slide, favoriscono il mantenimento dell'attività di *LCT*. Il DNA si ripiega e permette la comunicazione tra l'enhancer e il promotore, la regione dove si organizza l'avvio della lettura del gene.

La lattasi prodotta rimane la stessa: cambia la quantità che l'intestino continua a produrre. L'effetto sulla vita quotidiana è la capacità di digerire il lattosio anche da adulti.

A destra vediamo come la conoscenza della regolazione possa diventare una possibilità terapeutica.

BCL11A è un repressore: una proteina che contribuisce a ridurre la produzione di emoglobina fetale dopo la nascita. L'emoglobina fetale trasporta ossigeno durante la vita fetale e può essere utile anche in alcune malattie del sangue.

Con CRISPR possiamo modificare un enhancer di *BCL11A* nelle cellule staminali del sangue prelevate dal paziente. Quando queste cellule generano i precursori dei globuli rossi, producono meno repressore. Di conseguenza aumenta la produzione di emoglobina fetale.

Il passaggio da ricordare è questo: **riducendo un freno, permettiamo a un'altra istruzione di lavorare di più**. Questo principio è alla base di una terapia per pazienti selezionati con anemia falciforme o β-talassemia. Il trattamento richiede un percorso clinico complesso.

Questi esempi mostrano perché la regolazione è così importante: una variazione nel controllo di un gene può influenzare una capacità digestiva oppure offrire un modo per contrastare una malattia

Slide 21 Un solo testo moltissimi livelli di lettura

Riassumiamo le implicazioni: conoscere l'ordine delle lettere è fondamentale, ma dobbiamo anche capire come quel testo viene usato. Contano lo splicing, i regolatori, le modifiche chimiche e l'organizzazione tridimensionale. La complessità emerge dalle relazioni tra questi livelli. Ora cambiamo punto di osservazione e torniamo allo strumento con cui otteniamo il testo: come riusciamo a misurare le lettere di una molecola di DNA?

Slide 22 Leggere un genoma è diventato molto più accessibile

Le tecnologie hanno cambiato profondamente la scala della genomica. Qui confrontiamo due dati storici NHGRI: nel 2001 la stima era di circa 95 milioni di dollari per un genoma di dimensioni umane; nel maggio 2022 era di circa 525 dollari. Sono stime dei costi di produzione della sequenza secondo quella metodologia. Un percorso clinico comprende anche preparazione, analisi, interpretazione e refertazione. Con il passaggio da Sanger al sequenziamento massivo parallelo, e poi con le letture lunghe, sono cambiate anche le domande possibili: oggi possiamo confrontare grandi coorti e affrontare regioni prima molto difficili.

Slide 23 Evoluzione del sequenziamento del DNA

Questa immagine racconta come siamo passati dalla lettura manuale di piccoli tratti di DNA alla possibilità di leggere milioni di frammenti contemporaneamente.

Le lettere da riconoscere sono sempre le stesse — A, C, G e T — ma è cambiato il modo di leggerle. Seguiamo le tre tappe.

A sinistra troviamo il metodo Sanger manuale, sviluppato alla fine degli anni Settanta. Si parte da molte copie di un tratto di DNA e si utilizza una polimerasi per copiarlo. Nella reazione vengono inseriti anche nucleotidi speciali che, quando vengono incorporati, interrompono definitivamente la costruzione di quella copia.

Si ottiene così una raccolta di frammenti di lunghezze diverse. Nell'impostazione classica, le quattro reazioni vengono separate nelle corsie di un gel, indicate con A, C, G e T.

I frammenti più corti avanzano maggiormente nel gel. Per ricostruire la sequenza si leggono quindi le bande dal basso verso l'alto, annotando ogni volta la lettera della corsia corrispondente. Era un lavoro che richiedeva tempo, precisione e molta attenzione.

Al centro vediamo il Sanger automatizzato. Il principio rimane quello dell'interruzione della copia, ma le basi terminali vengono riconosciute attraverso marcatori fluorescenti. Negli strumenti a capillari, i frammenti vengono separati e passano davanti a un sistema ottico.

Il computer registra i segnali e produce il grafico con i picchi colorati che vediamo sullo schermo: l'elettroferogramma. Ogni picco contribuisce a identificare una lettera della sequenza.

L'automazione rende il lavoro più rapido e permette di analizzare più campioni. Il Sanger continua ancora oggi a essere utile per leggere regioni specifiche di DNA.

A destra arriviamo all'NGS, qui rappresentato dalla tecnologia Illumina. Il salto fondamentale è il parallelismo: sulla flow cell vengono letti contemporaneamente moltissimi frammenti, organizzati in cluster.

Il metodo è il sequencing by synthesis: la sequenza viene riconosciuta mentre la polimerasi costruisce una copia del DNA. A ogni ciclo si aggiunge una base, si registra il segnale e si prepara il ciclo successivo.

Qui c'è una differenza importante rispetto al Sanger: il blocco è temporaneo e reversibile, così la stessa copia può continuare ad allungarsi dopo ogni lettura.

Otteniamo tantissime sequenze corte, che il computer analizza e confronta con un genoma di riferimento oppure utilizza per ricostruire sequenze più lunghe.

Il messaggio della figura è quindi questo: la rivoluzione non è stata soltanto leggere più velocemente, ma leggere tantissimi frammenti insieme. Questo ha reso possibili studi su interi genomi, sull'attività dei geni e su grandi popolazioni: alla crescente capacità delle macchine si è affiancata una crescente necessità di analisi informatica.

Slide 24 COME SI PREPARA UNA LIBRERIA NGS?

Prima che un sequenziatore possa leggere il DNA, dobbiamo prepararlo. Il risultato si chiama libreria: una raccolta di frammenti di DNA ai quali aggiungiamo sequenze che permettono allo strumento di utilizzarli.

Possiamo immaginare di preparare tanti piccoli documenti per un lettore automatico. Ogni documento deve avere un formato compatibile e un'etichetta che ne indichi la provenienza.

La slide mostra un esempio di preparazione per il sequenziamento Illumina.

Nel primo passaggio, partiamo dal DNA già estratto e controllato e lo dividiamo in frammenti più piccoli. Il genoma è molto lungo: preparare frammenti permette di leggerne moltissimi in parallelo.

Nel secondo passaggio, alcuni enzimi preparano le estremità dei frammenti. Possiamo immaginarli come strumenti che sistemano i bordi dei documenti, rendendoli adatti al collegamento successivo.

Nel terzo passaggio, aggiungiamo gli adattatori. Sono brevi sequenze di DNA che permettono ai frammenti di interagire con la superficie del sequenziatore e con le molecole necessarie alla lettura.

Introduciamo anche gli indici, cioè codici identificativi dei campioni. Se mescoliamo librerie di persone diverse nella stessa corsa di sequenziamento, questi codici ci permettono poi di attribuire le letture al campione corretto. Sono come etichette sui bagagli: viaggiano insieme, ma possiamo riconoscere a chi appartengono.

Nel quarto passaggio, quando il protocollo lo prevede, usiamo la PCR per produrre più copie dei frammenti. Esistono anche protocolli chiamati PCR-free, nei quali questa amplificazione viene omessa. Le copie aumentano la quantità di materiale disponibile, senza aggiungere nuove informazioni al campione originale.

Nel quinto passaggio, puliamo la libreria e ne controlliamo quantità e lunghezza dei frammenti. Dobbiamo verificare che il materiale sia adatto alla lettura. Nella pratica, alcune pulizie avvengono anche tra i passaggi precedenti.

Infine, nel sesto passaggio, dosiamo le librerie e possiamo mescolare campioni con indici diversi. La miscela viene quindi preparata per il caricamento sul sequenziatore.

La preparazione della libreria precede la lettura del DNA. È una fase decisiva: la qualità di ciò che prepariamo influenza la qualità delle informazioni che otterremo.

Anche qui ritroviamo la complessità: per leggere un genoma servono molti passaggi coordinati, ognuno con una funzione precisa.

Slide 25 COME NASCE UN CLUSTER ILLUMINA

Prima di leggere il DNA, i sequenziatori Illumina devono prepararlo in modo che il suo segnale sia abbastanza intenso da essere riconosciuto. Il trucco è creare tante copie dello stesso frammento, tutte concentrate in un piccolo punto. Questo gruppo di copie si chiama cluster.

Possiamo immaginarlo come un coro: una sola voce, da lontano, è difficile da sentire; tante voci che cantano insieme la stessa nota producono un segnale molto più riconoscibile.

Seguiamo i quattro passaggi nella parte alta dell'immagine, che mostrano l'esempio classico dell'amplificazione a ponte.

Nel primo passaggio, il frammento si aggancia alla superficie. La flow cell è un piccolo dispositivo nel quale scorrono i reagenti. Sulla sua superficie sono fissate brevi sequenze di DNA, che funzionano come punti di aggancio. Gli adattatori aggiunti alle estremità dei nostri frammenti permettono di riconoscere questi punti e avviare la formazione di copie ancorate alla superficie.

Nel secondo passaggio si forma un ponte. Un filamento fissato alla superficie si piega e raggiunge un altro punto di aggancio vicino. Una DNA polimerasi, l'enzima che copia il DNA, costruisce il filamento complementare. Si forma così la struttura ad arco che vediamo nel disegno.

Nel terzo passaggio, i due filamenti si separano. Entrambi restano ancorati alla superficie e possono essere utilizzati come stampi per produrre altre copie.

Ripetendo il processo, si forma il cluster. Nello stesso piccolo spazio si accumulano molte copie derivate dal frammento iniziale. Contemporaneamente, in altri punti della flow cell, si formano cluster a partire da frammenti diversi.

Quando inizierà il sequenziamento, le copie di ciascun cluster verranno lette insieme. I loro segnali fluorescenti si sommeranno, permettendo allo strumento di riconoscere la base aggiunta in quel punto.

Guardiamo ora la parte bassa dell'immagine. Nelle flow cell non strutturate, i cluster si formano in posizioni non predefinite. Nelle flow cell patterned, invece, ci sono minuscoli pozzetti disposti secondo uno schema regolare: come posti numerati in un teatro. Questo aiuta a organizzare molti cluster e a distinguerne i segnali. Il meccanismo di amplificazione utilizzato su queste superfici può essere diverso dall'esempio classico a ponte mostrato sopra.

La frase da ricordare è: un cluster contiene tante copie dello stesso frammento; tanti cluster permettono di leggere contemporaneamente tanti frammenti diversi.

La formazione dei cluster prepara quindi il DNA alla lettura: è il passaggio che rende il suo segnale riconoscibile.

Slide 26 ILLUMINA SEQUENCING BY SYNTHESIS

Ora che abbiamo formato i cluster, possiamo iniziare a leggere il DNA. Il metodo si chiama sequencing by synthesis, cioè “sequenziamento mediante sintesi”: scopriamo la sequenza mentre ne costruiamo una copia.

La protagonista è la DNA polimerasi, l’enzima che aggiunge i mattoncini necessari a costruire un nuovo filamento. Per farlo segue la regola dell’appaiamento: A si associa a T e C a G. Se riconosciamo la base aggiunta alla copia, possiamo quindi risalire anche alla base presente sul filamento originale.

Seguiamo i tre passaggi dell’immagine.

Primo passaggio: aggiungiamo una base. Nella flow cell vengono introdotti i quattro tipi di nucleotide. La polimerasi incorpora quello complementare alla base dello stampo.

Questi nucleotidi hanno una caratteristica speciale: possiedono un blocco temporaneo, che impedisce di aggiungere subito il nucleotide successivo. È come scrivere una lettera e fermarsi, per dare alla macchina il tempo di leggerla. Per questo si chiamano terminatori reversibili: fermano la copia, ma il blocco può essere rimosso.

Secondo passaggio: leggiamo il segnale. Lo strumento registra la fluorescenza e riconosce quale base è stata aggiunta. I colori nell’immagine sono simbolici: le diverse piattaforme possono utilizzare codifiche ottiche differenti.

Ricordiamo che lo strumento osserva il segnale di un intero cluster, formato da molte copie dello stesso frammento. Queste copie lavorano insieme e producono un segnale abbastanza intenso da essere misurato.

Terzo passaggio: rimuoviamo il blocco. Vengono eliminati il segnale fluorescente utilizzato per la lettura e il blocco temporaneo. Il nuovo filamento può quindi allungarsi ancora.

A questo punto il ciclo ricomincia: aggiunta, lettura, rimozione del blocco. A ogni ciclo scopriamo un’altra lettera. Seguendo nel tempo i segnali dello stesso cluster, il computer ricostruisce la sequenza del frammento.

La parte bassa dell’immagine mostra il grande vantaggio: questo processo avviene contemporaneamente su moltissimi cluster, ciascuno associato a un frammento diverso.

Possiamo immaginare una classe enorme in cui ogni banco copia un testo differente: tutti scrivono una lettera, si fermano per il controllo e poi ripartono. La macchina segue tutti i banchi insieme.

La frase da ricordare è quella in fondo all’immagine: una base per ciclo, molti frammenti letti insieme. È questa combinazione di controllo e parallelismo che permette di produrre enormi quantità di sequenze.

Slide 27 I SEQUENZIATORI ILLUMINA

Questa immagine confronta tre sequenziatori Illumina. Usano lo stesso principio di lettura del DNA, ma possono produrre quantità di dati molto diverse. È come confrontare mezzi di trasporto con capacità differenti: la scelta dipende da quanto dobbiamo trasportare e dal viaggio che vogliamo fare.

Prima di guardare le macchine, chiariamo una parola: throughput significa quantità di dati prodotti in una corsa di sequenziamento.

Nell'immagine confrontiamo letture di tipo paired-end, indicate come "2 × 150 basi". Significa che ogni frammento viene letto dalle due estremità: 150 lettere da una parte e 150 dall'altra. Una coppia di letture fornisce quindi 300 basi di sequenza.

Partiamo dal MiSeq i100 Plus, a sinistra. Nella configurazione mostrata produce circa 100 milioni di coppie di letture, equivalenti a 30 miliardi di basi, cioè 30 gigabasi. È adatto, per esempio, a studiare genomi di microbi o virus, oppure gruppi selezionati di geni, chiamati pannelli mirati. Un genoma batterico è molto più piccolo del nostro: la stessa quantità di dati può quindi rispondere a domande molto diverse. Run time, output, and more

Al centro troviamo il NextSeq 2000. Con la flow cell P4 mostrata nell'immagine arriva a circa 1,8 miliardi di coppie di letture, cioè 540 gigabasi. Possiamo usarlo, per esempio, per studiare l'attività dei geni con l'RNA-seq, analizzare l'esoma — la parte del genoma che codifica le proteine — oppure leggere librerie ottenute da esperimenti su singole cellule. illumina.com

A destra troviamo il NovaSeq X Plus. Usando due flow cell 25B può produrre circa 16–21 terabasi in una corsa: migliaia di miliardi di lettere di DNA. Questa capacità permette di analizzare insieme molti campioni, come negli studi su grandi popolazioni. Per il sequenziamento dell'intero genoma umano, Illumina indica circa 128 genomi a copertura 30× nella configurazione mostrata. Capacity for high-intensity genomics

Ma perché diciamo 30×? Il genoma umano contiene circa tre miliardi di lettere, ma leggerne una sola copia non basta per ottenere un risultato affidabile. Cerchiamo di coprire ogni posizione, in media, con circa trenta letture. Alcune regioni riceveranno più letture, altre meno: questa ripetizione ci aiuta a riconoscere gli errori e le varianti.

I numeri di genomi nell'immagine sono quindi stime: assumono un budget di circa 120 gigabasi per campione, che tiene conto anche delle letture non utilizzabili. Con questo criterio, il NextSeq mostrato può ospitare circa quattro genomi umani, mentre una singola corsa del MiSeq mostrato non basta.

Il messaggio finale è quello scritto in basso: la domanda scientifica guida la scelta della macchina. Una maggiore capacità permette di studiare più campioni oppure di leggere più profondamente ciascun campione. Per rispondere bene, dobbiamo scegliere quanta informazione serve davvero.

Slide 28 Letture corte precise ma con un problema di contesto

Le letture corte sono molto utili perché permettono di produrre grandi quantità di dati con alta accuratezza. Spesso si leggono 150 basi da ciascuna estremità di un frammento. Il limite principale

riguarda il contesto: se un tratto ripetitivo è più lungo della lettura, possiamo non capire da quale copia provenga. Nell'immagine, i frammenti piccoli sono più difficili da ricollocare quando le frasi si assomigliano. Le short reads possono già identificare molte varianti, anche strutturali con metodi dedicati; per alcune ripetizioni o strutture complesse, però, una lettura lunga offre un vantaggio.

Slide 29 Oxford Nanopore sequencing

Possiamo immaginare il DNA come un lunghissimo filo sul quale è scritta una sequenza di lettere: A, C, G e T. Oxford Nanopore legge questo filo mentre attraversa un minuscolo poro, misurando un segnale elettrico.

Seguiamo il percorso mostrato nell'immagine.

Per prima cosa prepariamo il DNA. Cerchiamo di conservare frammenti lunghi, evitando di spezzarli troppo. Aggiungiamo degli adattatori che permettono di guidare la molecola verso il sistema di lettura.

Al centro della tecnologia c'è il nanoporo: una proteina con un'apertura piccolissima, inserita in una membrana. Attraverso questa apertura passa una corrente di ioni, cioè particelle cariche.

Una proteina motrice controlla l'avanzamento del DNA e ne accompagna un filamento attraverso il poro. Possiamo immaginarla come un meccanismo che fa scorrere lentamente un nastro, così che il lettore abbia il tempo di riconoscerne il contenuto.

Quando il DNA attraversa il poro, modifica la corrente elettrica. Il segnale dipende dal piccolo gruppo di basi che si trova in quel momento nella zona di lettura. Quindi lo strumento non fotografa le lettere del DNA: registra le variazioni della corrente.

Nell'immagine vediamo questo segnale come una linea che sale e scende, chiamata squiggle, cioè "scarabocchio". Ma come si passa da questa linea alle lettere?

Qui interviene il computer. Un programma, basato su reti neurali, riconosce le caratteristiche del segnale e le traduce in una sequenza di A, C, G e T. Questo passaggio si chiama basecalling.

Il grande vantaggio è che possiamo leggere un lungo tratto di DNA senza interromperci. La lunghezza ottenuta dipende anche da quanto è integra la molecola di partenza. Torna l'analogia del libro: leggere una frase intera aiuta a capire dove si trova, soprattutto quando nel testo ci sono molte parole ripetute.

Questa tecnologia permette anche di ricavare informazioni su alcune modifiche chimiche del DNA, come la metilazione, perché anch'esse possono influenzare il segnale elettrico. Inoltre, i dati diventano disponibili durante il sequenziamento: possiamo iniziare ad analizzarli mentre la lettura è ancora in corso.

La frase da ricordare è: con Oxford Nanopore, il passaggio del DNA attraverso un poro trasforma un'informazione molecolare in un segnale elettrico, che il computer traduce in sequenza.

Slide 30 PacBio HiFi leggere a lungo leggere più volte

Per capire la tecnologia PacBio, immaginiamo il genoma come un libro molto lungo, con intere frasi che si ripetono. Se possiamo leggere soltanto piccoli pezzi, a volte è difficile capire dove collocarli. Leggere frammenti più lunghi ci aiuta a ricostruire il testo, soprattutto nelle regioni ripetute.

Ma come funziona? Seguiamo i quattro passaggi dell'immagine.

Nel primo passaggio prepariamo il DNA. Alle due estremità di un lungo frammento aggiungiamo piccoli adattatori a forcina. Questi collegano i due filamenti e creano un percorso chiuso, chiamato SMRTbell. Il vantaggio è che la macchina potrà leggere più volte lo stesso frammento.

Nel secondo passaggio osserviamo una singola molecola. Il DNA viene associato a una DNA polimerasi, un enzima capace di copiarlo, simile a quelli che lavorano nelle nostre cellule. La polimerasi si trova sul fondo di un minuscolo pozzetto, progettato per osservare ciò che accade proprio vicino all'enzima.

Nel terzo passaggio leggiamo i segnali luminosi. Per costruire la copia, la polimerasi aggiunge uno dopo l'altro i nucleotidi, cioè i mattoncini del DNA. Nella reazione, i quattro tipi di nucleotide portano marcatori fluorescenti distinguibili. Quando vengono incorporati, producono segnali che lo strumento registra e traduce nelle lettere A, C, G e T.

È come osservare qualcuno mentre copia un testo e riconoscere ogni lettera dal segnale luminoso che accompagna il suo inserimento.

Il quarto passaggio è il segreto delle letture HiFi. Grazie al percorso circolare, la polimerasi può attraversare più volte lo stesso frammento, leggendo entrambi i filamenti. Il computer confronta questi passaggi e costruisce una sequenza di consenso: gli errori occasionali possono così essere corretti. HiFi significa proprio "alta fedeltà".

Nella parte bassa dell'immagine vediamo il vantaggio delle letture lunghe: possono attraversare una regione ripetuta e raggiungere le sequenze diverse che si trovano ai suoi lati. Queste funzionano come punti di riferimento e aiutano a collocare correttamente il frammento nel genoma.

La frase da ricordare è quindi: PacBio HiFi combina due vantaggi: leggere frammenti lunghi e leggerli più volte per ottenere una sequenza molto accurata. Questo ci aiuta a esplorare parti del genoma difficili da ricostruire e a riconoscere cambiamenti anche di grandi dimensioni.»

Slide 31 Quale tecnologia Dipende dalla domanda

Questa tabella serve a scegliere in base alla domanda. Se vogliamo analizzare molte varianti puntiformi in grandi coorti, le short reads possono essere molto efficienti. Se dobbiamo ricostruire un genoma o risolvere una regione ripetitiva, le letture lunghe possono offrire informazioni decisive. PacBio HiFi punta su un consenso molto accurato; Nanopore offre letture elettriche continue e, in alcune condizioni, estremamente lunghe. Accuratezza e costi dipendono dalla preparazione, dalla piattaforma e dall'analisi. In molti progetti si combinano approcci diversi, perché informazioni complementari possono rispondere meglio alla stessa domanda biologica.

Slide 32 Una variante non è automaticamente una malattia

Quando confrontiamo un genoma con un riferimento troviamo varianti di dimensioni diverse. Può cambiare una sola base; poche basi possono essere inserite o eliminate; un segmento più grande può essere duplicato, invertito o spostato. In altri casi aumenta il numero di ripetizioni. La parola variante descrive una differenza, non una diagnosi. Molte differenze sono innocue; alcune modificano un rischio o hanno effetti protettivi; altre possono causare malattia. Per interpretarle dobbiamo considerare frequenza nella popolazione, funzione, ereditarietà e quadro clinico. Vediamo ora tre esempi in cui il tipo di cambiamento è particolarmente istruttivo.

Slide 33 Anemia falciforme una base può contare molto

Nell'anemia falciforme, una variante del gene HBB sostituisce una A con una T e cambia un amminoacido dell'emoglobina. Si forma l'emoglobina S, che in certe condizioni può aggregarsi e deformare i globuli rossi. Questo contribuisce ad anemia e problemi nella circolazione. È una malattia monogenica, ma la sua gravità dipende anche da altri fattori. La terapia exa-cel usa CRISPR per modificare un enhancer di BCL11A nelle cellule staminali del paziente e aumentare l'emoglobina fetale. Non corregge direttamente HBB. È un trattamento specialistico per pazienti selezionati, con un percorso complesso che comprende preparazione, modifica delle cellule e reinfusione.

Slide 34 Fibrosi cistica stesso gene molti effetti

Nella fibrosi cistica il gene coinvolto è CFTR, che codifica un canale per il trasporto di ioni. Se il canale funziona male, cambiano il movimento degli ioni e dell'acqua e le secrezioni possono diventare più dense. Sono descritte migliaia di varianti, con effetti e livelli di evidenza differenti. Nell'esempio F508del manca un amminoacido della proteina. Alcune terapie modulatrici aiutano specifiche forme alterate di CFTR a funzionare meglio. Per questo conoscere il genotipo può orientare la scelta terapeutica, insieme all'età, alle indicazioni del farmaco e alla valutazione clinica. Lo stesso nome di malattia può quindi comprendere situazioni molecolari diverse.

Slide 35 Huntington quando una parola si ripete troppo

Nella malattia di Huntington il cambiamento riguarda il numero di ripetizioni CAG nel gene HTT. Il tratto ripetuto può espandersi oltre determinate soglie. Fra 36 e 39 ripetizioni non tutte le persone sviluppano la malattia; con 40 o più la probabilità nell'arco della vita è molto elevata. Il numero, però, non permette di prevedere con certezza l'età d'esordio. Le espansioni mostrano perché la lunghezza della lettura può contare: un tratto ripetuto lungo può essere difficile da ricostruire con letture corte. Esistono anche test mirati consolidati. Un eventuale test predittivo richiede un percorso di consulenza genetica.

Slide 36 Malattie complesse molti piccoli effetti

Per molte malattie frequenti non troviamo una sola variante capace di spiegare tutto. Nel diabete di tipo 2, in molte malattie cardiovascolari e in vari disturbi psichiatrici contribuiscono numerose varianti, spesso di piccolo effetto, insieme ad ambiente e storia personale. Gli studi GWAS cercano associazioni fra varianti e caratteristiche misurate in molte persone. Ogni picco del grafico indica una regione associata, non necessariamente il gene responsabile. Un punteggio di rischio poligenico combina molti effetti stimati. Può contribuire a una valutazione del rischio, ma richiede una validazione adatta alla popolazione e non determina il destino di una persona.

Slide 37 QUANDO IL GENOMA RIVELA LA CAUSA DI UNA MALATTIA

Finora abbiamo visto come si legge il DNA. Ora vediamo che cosa può significare questa lettura per una persona che cerca una spiegazione alla propria malattia.

Il primo caso riguarda due gemelli di 14 anni con disturbi del movimento. Il sequenziamento dell'intero genoma ha identificato alterazioni nelle due copie di un gene chiamato SPR. Questo gene contribuisce alla produzione di sostanze necessarie per ottenere alcuni messaggeri chimici del cervello, tra cui dopamina e serotonina.

Conoscere il difetto preciso ha permesso ai medici di adattare una terapia già in corso, ottenendo un miglioramento dei sintomi. È un esempio concreto di come una diagnosi genetica possa aiutare a scegliere una cura più adatta.

Il secondo caso riguarda una ragazza con sindrome di Alport, una malattia che può danneggiare i reni. Un esame precedente, l'esoma, non aveva identificato la causa genetica. L'esoma legge soprattutto le parti dei geni che contengono le istruzioni per costruire le proteine.

Il sequenziamento dell'intero genoma ha invece trovato una variante in un introne, una regione normalmente rimossa durante il montaggio dell'RNA. Quella modifica faceva includere nel messaggio un tratto che avrebbe dovuto essere eliminato: come inserire una frase estranea nelle istruzioni di montaggio. Il risultato era un messaggio alterato.

Questo ci ricorda che anche il DNA che non codifica direttamente una proteina può avere conseguenze importanti.

Il terzo esempio è la sindrome CANVAS, che provoca problemi di equilibrio, coordinazione e sensibilità. In questo caso la causa individuata era una breve sequenza di DNA ripetuta troppe volte, in entrambe le copie del gene RFC1. Immaginate una parola copiata centinaia di volte all'interno di un testo: il problema non è necessariamente una lettera sbagliata, ma il numero di ripetizioni.

Le analisi del genoma hanno fornito gli indizi, poi verificati con altri esami, per spiegare casi rimasti senza diagnosi.

Questi tre esempi mostrano diversi aspetti della complessità: può essere alterata un'istruzione, il montaggio del messaggio oppure la lunghezza di una regione ripetuta. Trovare la causa non significa sempre trovare subito una cura, ma può dare una spiegazione, orientare l'assistenza e, in alcuni casi, cambiare il trattamento.»

Slide 38 LONG READS: CAUSE GENETICHE PRIMA NASCOSTE

La slide precedente ci ha mostrato perché studiare il genoma può essere utile. Qui aggiungiamo una domanda: conta anche il modo in cui lo leggiamo?

Immaginate di ricostruire un libro partendo da piccoli frammenti. Se molte frasi si assomigliano, oppure mancano dei pezzi, capire dove collocare ogni frammento diventa difficile. Le letture lunghe di PacBio e Oxford Nanopore permettono di osservare tratti più estesi di DNA e di riconoscere meglio alcune di queste situazioni.

Nel primo caso, un paziente con complesso di Carney, una rara sindrome, aveva già effettuato diversi esami, compreso il sequenziamento del genoma con letture corte. La causa genetica era però rimasta sconosciuta.

Con PacBio è stata individuata una delezione: mancavano 2.184 lettere di DNA in una regione del gene PRKAR1A. Tornando al libro, le letture lunghe hanno aiutato a riconoscere una porzione mancante, chiarendo la causa della malattia.

Il secondo caso riguarda la sindrome di Alström, che può interessare vista, udito e metabolismo. Per questa malattia recessiva, il difetto deve coinvolgere entrambe le copie del gene ALMS1, quella ereditata dalla madre e quella ereditata dal padre.

L'esoma aveva trovato una sola delle due varianti. Il sequenziamento mirato Oxford Nanopore ha identificato l'altra: un elemento di DNA chiamato Alu si era inserito nel gene, interrompendone la sequenza. È come trovare un paragrafo estraneo dentro un'istruzione. In questo caso il precedente esame era un esoma, quindi anche la diversa estensione dell'analisi ha contribuito alla scoperta.

Il terzo esempio è l'ipoventilazione centrale congenita, una malattia che compromette il controllo automatico della respirazione, soprattutto durante il sonno. PacBio HiFi ha individuato nel gene PHOX2B una ripetizione con 18 basi aggiuntive, sfuggita agli esami precedenti. Il risultato è stato poi confermato con un test clinico.

Una precisazione: questi casi non dimostrano che le short reads siano sempre incapaci di trovare tali alterazioni. Contano anche la preparazione del campione e i programmi di analisi; alcune varianti possono emergere rianalizzando i dati.

Il collegamento con la complessità è questo: per capire il genoma dobbiamo leggere le lettere, ma anche ricostruire come sono organizzate. Pezzi mancanti, inserzioni e ripetizioni possono cambiare il significato delle istruzioni. Le long reads ci aiutano a vedere meglio questa organizzazione.

Slide 39 Dalla complessità a diagnosi e terapie più precise

Siamo arrivati agli effetti della complessità in medicina. Una tecnologia ci permette di leggere una differenza; le conoscenze biologiche e cliniche aiutano a stabilire se quella differenza sia rilevante. Da qui possono nascere una diagnosi o una terapia più precisa. Per interpretare un singolo caso, però, dobbiamo sapere anche quanto una variante sia frequente e in quali contesti compaia. Il passo successivo è quindi osservare molte persone insieme.

Slide 40 Per capire una variante una persona non basta

Immaginiamo di trovare una variante in una persona con una malattia. Come capiamo se c'è un legame? Un confronto con molti altri genomi ci aiuta a sapere se quella variante è frequente o rara e se compare associata a un tratto. Una variante rara non è automaticamente patogena. I grandi cataloghi e le biobanche permettono di collegare dati genetici, salute e altre informazioni. Gli studi devono comunque considerare differenze tra gruppi, possibili fattori confondenti e qualità dei dati. Dall'osservazione collettiva ricaviamo indizi che poi vanno verificati e possono aiutare anche il singolo paziente.

Slide 41 EUROPA: GRANDI STUDI SUL DNA DELLE PERSONE

Perché studiare il genoma di centinaia di migliaia di persone? Perché il significato di una variante diventa più chiaro quando possiamo confrontarla con quelle di molte altre persone e collegarla a informazioni sulla salute.

Immaginiamo di trovare una parola insolita in un libro: potrebbe essere un errore oppure una parola corretta che non conosciamo. Consultare una grande biblioteca ci aiuta a capirlo. Le raccolte genomiche svolgono una funzione simile.

La mappa mostra alcune importanti iniziative europee. UK Biobank associa i genomi di circa mezzo milione di persone a informazioni sulla salute e sullo stile di vita. Seguendo queste persone nel tempo, possiamo studiare come genetica e ambiente contribuiscano al rischio di malattia.

In Islanda, deCODE studia la variazione genetica insieme a informazioni sanitarie e familiari. In Finlandia, FinnGen collega dati genetici e registri sanitari. Queste risorse aiutano anche a individuare varianti che, per la storia di una popolazione, possono essere più frequenti e quindi più facili da studiare.

Genomics England parte soprattutto da domande cliniche, come la ricerca delle cause di malattie rare e lo studio dei tumori. La biobanca estone e Our Future Health contribuiscono invece a comprendere le relazioni fra patrimonio genetico, salute e prevenzione.

Attenzione alla tabella: numero di partecipanti e numero di genomi interamente sequenziati non sono la stessa cosa. Alcuni progetti analizzano marcatori, cioè posizioni selezionate del DNA.

Nella parte bassa vediamo come l'Europa cerca di collegare queste risorse. Genome of Europe, al quale partecipa anche l'Italia, punta a costruire una raccolta di almeno 100.000 genomi di riferimento. L'iniziativa 1+ Million Genomes e l'infrastruttura GDI lavorano per consentire un accesso sicuro ai dati fra paesi, rispettando consenso e privacy.

Il collegamento con la complessità è questo: per comprendere una persona dobbiamo studiare molte persone, mantenendo insieme informazioni genetiche, cliniche e ambientali

Slide 42 IL MONDO: MILIONI DI GENOMI, DIVERSITÀ DA INCLUDERE

Allargando lo sguardo al mondo, vediamo che la genomica sta diventando un'impresa collettiva su scala enorme.

Negli Stati Uniti, All of Us raccoglie dati genomici e sanitari dando particolare attenzione a comunità storicamente poco rappresentate nella ricerca. Il Million Veteran Program studia la salute dei veterani, mentre TOPMed approfondisce soprattutto malattie del cuore, dei polmoni e del sangue, insieme ai disturbi del sonno.

In altre regioni troviamo grandi programmi come quello degli Emirati, PRECISE a Singapore e la Tohoku Medical Megabank in Giappone. Le iniziative in India, Cina, Brasile e Qatar contribuiscono a descrivere una diversità genetica che le raccolte europee, da sole, non potrebbero rappresentare.

A che cosa serve questa diversità? Immaginiamo che una variante sembri rarissima in una banca dati composta soprattutto da persone di origine europea. Se quella variante è comune in un'altra popolazione poco studiata, potremmo interpretarla male. Conoscere meglio le diverse popolazioni aiuta a distinguere la normale variabilità umana dalle varianti che meritano ulteriori indagini.

Questi studi possono anche indicare nuove possibilità terapeutiche. Un esempio è PCSK9: lo studio di persone con varianti che riducono l'attività di questo gene ha aiutato a comprendere come abbassare il colesterolo e ha contribuito allo sviluppo di farmaci diretti contro quella proteina.

I colori della mappa richiedono attenzione. Il verde indica iniziative per cui riportiamo dati già ottenuti; l'oro evidenzia programmi per cui mostriamo un obiettivo futuro. Il grigio significa che il paese non è incluso nella nostra selezione: non che sia privo di ricerca.

Molte popolazioni africane, latinoamericane e indigene restano comunque poco rappresentate. Questo limita le conoscenze e può ridurre l'affidabilità di alcuni strumenti di previsione del rischio.

Per ottenere una medicina più precisa per tutti, servono grandi numeri e una partecipazione più rappresentativa dell'umanità. La complessità comprende anche la nostra storia, i luoghi in cui viviamo e le condizioni della nostra vita

Slide 43 Chi è rappresentato nei dati È una domanda di equità

Chi partecipa agli studi influenza ciò che possiamo imparare. Nei campioni aggregati dei GWAS mostrati qui, circa l'88 per cento della rappresentazione riguarda ascendenza europea. Questo conteggio riguarda partecipazioni agli studi, con possibili riusi delle stesse coorti, non individui unici nel mondo. Se un modello viene costruito soprattutto su alcuni gruppi, può stimare peggio il rischio in altri. L'ascendenza descrive relazioni genetiche e storia delle popolazioni, non categorie umane rigide. Ampliare la diversità significa migliorare le conoscenze, coinvolgendo le comunità in modo rispettoso e condividendo i benefici della ricerca.

Slide 44 Scoperte 1 Il genoma racconta incontri e adattamenti

I genomi raccontano anche incontri tra popolazioni del passato. Gli studi sul DNA antico, riconosciuti con il Nobel 2022 a Svante Pääbo, hanno mostrato contributi di ominini estinti nei genomi umani attuali. Molte persone con ascendenza fuori dall'Africa hanno circa l'uno o due per cento di DNA neanderthaliano, con una storia più articolata di una semplice divisione fra continenti. In Tibet, un aplotipo di EPAS1 di origine denisoviana contribuisce all'adattamento all'alta quota. La persistenza della lattasi, invece, è emersa attraverso varianti diverse in più popolazioni pastorali. Genetica, ambiente e cultura si sono influenzati nel tempo.

Slide 45 Milioni di differenze dentro un patrimonio condiviso

Ognuno di noi presenta milioni di varianti rispetto a un riferimento: indicativamente quattro o cinque milioni di siti, a cui si aggiunge la complessità dei cambiamenti strutturali. Allo stesso tempo condividiamo la grande maggioranza della sequenza. Non c'è quindi un genoma individuale perfetto, privo di differenze, che rappresenti tutte le persone. Anche la percentuale esatta di somiglianza dipende da quali cambiamenti contiamo. Il punto biologico è che una variante acquista significato attraverso le sue relazioni con il resto del genoma, con le funzioni cellulari e con l'ambiente.

Slide 46 STRUMENTI PER RISOLVERE LA COMPLESSITA'

TECNOLOGIE: Quali strumenti tecnologici servono per studiare la complessità del genoma?

COMPETENZE: Quali competenze servono per studiare la complessità del genoma e del trascrittoma?

ACCESSO: come l'accesso alle tecnologie può rendere la ricerca competitiva.

Slide 47 - 48 e 49 In Italia la Piattaforma Nazionale di Genomica

Come rendiamo possibili questi studi anche in Italia? La Piattaforma Nazionale di Genomica di Human Technopole mette a disposizione della comunità di ricerca tecnologie e competenze specialistiche. Il lavoro comprende preparazione dei campioni, sequenziamento short e long read e supporto all'analisi, secondo i servizi e le procedure disponibili. L'accesso avviene attraverso bandi e percorsi delle Piattaforme Nazionali. Possono partecipare anche ricercatori che lavorano su domande cliniche: questo significa sostenere la ricerca collegata alla salute, non offrire indistintamente un servizio di refertazione diagnostica. L'obiettivo è permettere a buone domande scientifiche di incontrare gli strumenti necessari per affrontarle.

Slide 50 WHOLE GENOME SEQUENCING

Questa immagine mostra come possiamo studiare il DNA di una persona attraverso il Whole Genome Sequencing, cioè il sequenziamento dell'intero genoma.

Immaginiamo il genoma come un'enciclopedia di circa tre miliardi di lettere. Con le short reads Illumina non la leggiamo dall'inizio alla fine in un unico passaggio: leggiamo tantissimi piccoli frammenti e utilizziamo il computer per capire dove si trovano.

Seguiamo i sei passaggi dell'immagine.

Primo: estraiamo il DNA. Partiamo da un campione biologico, per esempio sangue o saliva, dal quale recuperiamo il DNA contenuto nelle cellule. Questo DNA sarà il materiale da analizzare.

Secondo: prepariamo la libreria. Il DNA viene frammentato e alle estremità dei frammenti vengono aggiunti degli adattatori. La parola “libreria” indica proprio questa raccolta di frammenti preparati per il sequenziamento. Gli adattatori consentono di agganciarli alla superficie di lettura e avviare le reazioni.

Terzo: formiamo i cluster. Sulla flow cell vengono prodotte molte copie di ciascun frammento, concentrate nello stesso punto. Perché? Il segnale di una singola copia sarebbe troppo debole: tante copie che lavorano insieme producono un segnale riconoscibile.

Quarto: leggiamo il DNA attraverso la luce. La polimerasi costruisce una copia di ciascun frammento aggiungendo una base per ciclo. Lo strumento registra i segnali fluorescenti e il computer li traduce nelle lettere A, C, G e T. Questo processo avviene su moltissimi cluster contemporaneamente.

Otteniamo così un’enorme raccolta di letture corte, spesso di 150 basi da ciascuna estremità del frammento.

Quinto: collochiamo le letture. Il computer confronta questi piccoli testi con un genoma di riferimento, che funziona come una mappa. Cerca la posizione più plausibile per ogni lettura: questo passaggio si chiama allineamento.

Nella parte bassa dell’immagine compare il termine copertura 30×: significa che ogni posizione viene letta circa trenta volte, in media. Alcune saranno lette di più, altre di meno. Avere più osservazioni ci aiuta a distinguere una differenza reale da un errore di lettura.

Sesto: cerchiamo le varianti. Nell’esempio, il riferimento presenta una G, mentre le letture mostrano una T. Questa differenza è una possibile variante da valutare.

Attenzione però: il riferimento è una mappa, non un modello di persona “perfetta”. Le differenze fanno parte della diversità umana e una variante non indica automaticamente una malattia.

Il Whole Genome Sequencing permette di esplorare sia i geni sia le regioni non codificanti. Tuttavia, con le letture corte alcune zone, soprattutto quelle molto ripetitive, possono restare difficili da analizzare.

La frase da ricordare è: tante letture corte, analizzate insieme, ci danno una visione dell’intero genoma. Leggere le lettere è il primo passo; comprenderne il significato è la sfida successiva.

Slide 51 RNA SEQUENCING

Con il sequenziamento del genoma cerchiamo di leggere le istruzioni contenute nel DNA. Con l’**RNA sequencing**, o *RNA-seq*, ci poniamo una domanda diversa: **quali di queste istruzioni vengono utilizzate nel campione che stiamo studiando?**

Possiamo immaginare il DNA come una biblioteca e gli RNA messaggeri come le copie delle pagine che le cellule stanno utilizzando. Cellule diverse possiedono sostanzialmente la stessa biblioteca, ma consultano pagine differenti.

Seguiamo i sei passaggi dell’immagine.

Primo: raccogliamo l'RNA. Partiamo da un campione che contiene molte cellule ed estraiamo le molecole di RNA. La loro abbondanza dipende dal tipo di cellule, dalle condizioni in cui si trovano e dal momento in cui raccogliamo il campione.

Secondo: selezioniamo i messaggi che ci interessano. Nell'esempio dell'immagine arricchiamo gli RNA messaggeri, o *mRNA*, che portano le informazioni per produrre proteine. Riduciamo invece la presenza degli RNA ribosomiali, molto abbondanti, che altrimenti occuperebbero gran parte delle letture. Esistono anche protocolli che permettono di studiare una gamma più ampia di RNA.

Terzo: copiamo l'RNA in DNA. Nel metodo Illumina illustrato non leggiamo direttamente l'RNA. Un enzima chiamato *trascrittasi inversa* ne produce una copia in DNA, detta **DNA complementare**, o *cDNA*. Da questo materiale prepariamo una libreria di frammenti con gli adattatori necessari al sequenziamento.

Quarto: leggiamo il cDNA. Si formano i cluster e si utilizza il sequenziamento per sintesi: una base per ciclo, moltissimi frammenti letti contemporaneamente. Otteniamo così tante brevi sequenze che rappresentano gli RNA presenti nel campione.

Quinto: assegniamo le letture. Il computer cerca di capire da quali geni provengono. Nell'immagine, il gene A riceve molte letture, il gene B un numero intermedio e il gene C poche.

In generale, più molecole di RNA di un gene sono presenti, più letture possiamo ottenere da quel gene. Ma i conteggi richiedono correzioni: dipendono anche dalla quantità totale di dati e dalle caratteristiche dei trascritti.

Sesto: confrontiamo i campioni. Dopo la normalizzazione, possiamo chiederci quali geni mostrino un'espressione diversa, per esempio prima e dopo un trattamento. Nel grafico illustrativo, il gene A diminuisce, il gene B aumenta e il gene C rimane simile.

Ricordiamo però un limite importante: questo è **bulk RNA-seq**, cioè l'analisi congiunta di molte cellule. Il risultato descrive il campione nel suo insieme. Una differenza può dipendere sia da cambiamenti nelle cellule sia dalla presenza di proporzioni diverse di tipi cellulari.

La frase da ricordare è: **l'RNA-seq ci offre una fotografia dei messaggi presenti in un campione in un dato momento.** Ci aiuta a capire come viene utilizzato il genoma, anche se da solo non ci dice quante proteine siano effettivamente prodotte

Slide 52 SINGLE CELL RNA SEQUENCING

Un tessuto può sembrare uniforme, ma contiene cellule diverse, che svolgono compiti differenti. **Il single-cell RNA sequencing ci permette di studiare i messaggi prodotti dalle singole cellule, distinguendo differenze che nell'analisi del tessuto intero potrebbero restare nascoste.**

Pensiamo a un'orchestra: ascoltarla nel suo insieme ci fa riconoscere il brano, ma distinguere i singoli strumenti ci aiuta a capire chi sta suonando e quale contributo sta dando. Qui facciamo qualcosa di simile con le cellule.

Seguiamo i cinque passaggi dell'immagine.

Primo: partiamo da un tessuto. I colori rappresentano cellule con caratteristiche diverse. Per esempio, un tessuto può contenere cellule immunitarie, cellule dei vasi sanguigni e cellule specializzate dell'organo.

Secondo: separiamo le cellule. Il tessuto viene trattato per ottenere una sospensione di cellule singole. È un passaggio delicato: dobbiamo cercare di preservare sia le cellule sia l'RNA che contengono.

Terzo: assegniamo un'etichetta a ogni cellula. Nell'approccio illustrato, le cellule vengono distribuite in minuscole goccioline insieme a microsfere che portano brevi sequenze identificative, chiamate *barcode*.

Durante la preparazione, questi codici vengono incorporati nelle copie dei messaggi RNA. I messaggi provenienti dalla stessa cellula ricevono lo stesso identificativo. Non è un vero codice a barre stampato: è una sequenza di lettere del DNA che il computer saprà riconoscere.

Questo è il passaggio decisivo: **possiamo mescolare il materiale e leggerlo insieme, conservando l'informazione sulla cellula di provenienza**.

Quarto: leggiamo i messaggi. L'RNA viene copiato in DNA complementare, o *cDNA*, e preparato per il sequenziamento. Le letture ci permettono di identificare i geni da cui provengono i messaggi; i barcode permettono di assegnarli alle cellule.

Quinto: confrontiamo le cellule. Il computer costruisce per ciascuna un profilo di espressione: quali geni sono rappresentati e quanto RNA osserviamo per ciascuno.

Nel grafico finale ogni punto rappresenta una cellula. Punti vicini indicano profili simili: possiamo così riconoscere tipi cellulari diversi oppure stati differenti dello stesso tipo cellulare, per esempio cellule immunitarie a riposo o attivate. **Questa è una mappa di somiglianza, non della posizione delle cellule nel tessuto.**

Perché è utile? In un'infiammazione, per esempio, possiamo capire quali popolazioni cellulari rispondono e quali messaggi producono. Possiamo anche individuare cellule rare, il cui segnale sarebbe poco visibile nella media del campione.

La fotografia rimane parziale: non rileviamo ogni molecola di RNA e osserviamo le cellule in un determinato momento.

La frase da ricordare è quella in basso: **"Stesso DNA, geni usati in modo diverso".** La complessità di un tessuto nasce anche dalla diversità delle cellule che lo compongono e dal modo in cui collaborano

Slide 53 TRASCRITTOMICA SPAZIALE

Con il single-cell RNA sequencing abbiamo imparato a distinguere i messaggi prodotti dalle singole cellule. Ma quando separiamo le cellule di un tessuto, generalmente perdiamo un'informazione importante: **dove si trovavano e quali erano le loro vicine**.

La **trascrittomica spaziale** aggiunge proprio questa dimensione. Ci aiuta a rispondere a due domande insieme: **quali geni sono attivi e dove osserviamo questa attività nel tessuto?**

Pensiamo a una città: conoscere le professioni degli abitanti è utile, ma sapere dove vivono e lavorano ci aiuta a capire come funziona la città. Anche nei tessuti, la posizione delle cellule e il loro vicinato contano.

Seguiamo i quattro passaggi dell'immagine.

Primo: prepariamo una sezione di tessuto. Una fettina sottile viene collocata su un supporto. Conserviamo così l'organizzazione del tessuto: possiamo riconoscere le sue regioni e osservare come sono disposte le cellule.

Secondo: rileviamo i messaggi RNA. Come nell'RNA-seq, studiamo le molecole di RNA per ottenere informazioni sull'espressione dei geni. Nell'immagine i colori rappresentano messaggi diversi: alcuni sono più abbondanti in una zona, altri in un'altra.

Esistono tecnologie differenti: alcune catturano gli RNA su un supporto e li analizzano mediante sequenziamento; altre riconoscono i messaggi direttamente nel tessuto attraverso segnali osservabili al microscopio.

Terzo: associamo un indirizzo. Questo è il passaggio fondamentale. Ogni segnale viene collegato a una posizione. Nei metodi basati su cattura e sequenziamento, particolari codici molecolari possono identificare il punto del supporto dal quale proviene il messaggio. È come aggiungere un indirizzo a una lettera.

Quarto: costruiamo una mappa. Il computer combina le informazioni sull'RNA con l'immagine del tessuto. Possiamo così vedere dove si concentrano i messaggi di determinati geni e confrontare regioni differenti.

Per esempio, in un tessuto infiammato possiamo chiederci dove si trovano le cellule immunitarie e quali programmi di espressione osserviamo nelle cellule vicine. Nel cervello possiamo studiare come questi programmi cambiano tra strati e regioni.

La vicinanza, da sola, non dimostra che due cellule stiano comunicando, ma ci aiuta a formulare ipotesi da verificare.

Attenzione anche alla **risoluzione**: a seconda della tecnologia, un punto di misura può comprendere più cellule oppure permettere di distinguere dettagli all'interno di una cellula.

La frase da ricordare è quella in fondo all'immagine: **"Non solo quali geni sono attivi: anche dove".** La complessità di un tessuto dipende dalle cellule che lo compongono, ma anche da come sono organizzate e dalle relazioni che possono stabilire

Slide 54 Perché l'accesso conta quanto l'innovazione

Gli strumenti sono costosi e richiedono competenze che si costruiscono nel tempo. Condividerli consente a più gruppi di ricerca di lavorare su domande importanti. Serve però un'intera organizzazione: personale formato, qualità, standard, capacità di calcolo e interpretazione. Per trasferire un metodo dalla ricerca alla clinica occorrono anche validazione e responsabilità definite. I dati genomici possono identificare una persona e riguardare i suoi familiari; consenso e tutela dei dati fanno parte del percorso. L'accessibilità delle tecnologie comprende quindi sia la possibilità di usarle sia le condizioni per produrre risultati affidabili e utili.

Slide 56 Tre idee da portare a casa

Torniamo alle tre parole iniziali. I presupposti sono un genoma lungo, ripetitivo e organizzato, che oggi possiamo studiare anche in regioni prima inaccessibili. Le implicazioni sono le tecnologie e i livelli di regolazione necessari per leggerlo e comprenderlo. Gli effetti sono le scoperte sulla nostra storia e gli strumenti che possono migliorare diagnosi e terapie. Le risposte più utili arrivano quando colleghiamo questi livelli. Studiare la complessità significa riconoscere le relazioni che rendono importante una sequenza in una determinata cellula, persona o popolazione.

Slide 57 Grazie

Grazie per l'attenzione e per le vostre domande. Spero che questo viaggio abbia mostrato come quattro lettere possano partecipare a un sistema ricco di relazioni, dalla struttura del DNA alla salute delle persone e alla storia delle popolazioni. Le fonti riportate nella presentazione permettono di approfondire i singoli argomenti. Se ripensate alla domanda iniziale, forse ora vedete quanti passaggi collegano una sequenza a una persona. Sono proprio questi passaggi che continuiamo a studiare.